W604–W608 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 14 June 2012*
doi:10.1093/nar/gks562

RegPrecise web services interface: programmatic access to the transcriptional regulatory interactions in bacteria reconstructed by comparative genomics

Pavel S. Novichkov[1,*], Thomas S. Brettin[2], Elena S. Novichkova[1], Paramvir S. Dehal[1], Adam P. Arkin[1], Inna Dubchak[1] and Dmitry A. Rodionov[3,4,*]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6420, USA, [3]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994 and [4]Sanford-Burnham Medical Research Institute, La Jolla, CA 92037, USA

Received March 19, 2012; Revised May 10, 2012; Accepted May 18, 2012

ABSTRACT

Web services application programming interface (API) was developed to provide a programmatic access to the regulatory interactions accumulated in the RegPrecise database (http://regprecise.lbl.gov), a core resource on transcriptional regulation for the microbial domain of the Department of Energy (DOE) Systems Biology Knowledgebase. RegPrecise captures and visualize regulogs, sets of genes controlled by orthologous regulators in several closely related bacterial genomes, that were reconstructed by comparative genomics. The current release of RegPrecise 2.0 includes >1400 regulogs controlled either by protein transcription factors or by conserved ribonucleic acid regulatory motifs in >250 genomes from 24 taxonomic groups of bacteria. The reference regulons accumulated in RegPrecise can serve as a basis for automatic annotation of regulatory interactions in newly sequenced genomes. The developed API provides an efficient access to the RegPrecise data by a comprehensive set of 14 web service resources. The RegPrecise web services API is freely accessible at http://regprecise.lbl.gov/RegPrecise/services.jsp with no login requirements.

INTRODUCTION

Genome-scale transcriptional regulatory network (TRN) for any specific microbial organism is a critical component on the way to the next big milestone in System Biology—building an integrated metabolic and regulatory model that can accurately predict cellular growth phenotypes.

Several approaches and associated web resources have been developed for genome-wide reconstruction of metabolic pathways for a number of microbial genomes. BioCyc maintains an encyclopedia of experimentally defined metabolic pathways in model organisms and uses it for reconstruction of metabolism in other sequenced genomes (1). Model SEED allows user to submit any complete genome to the annotation pipeline to generate a draft metabolic model (2).

At the same time, the large-scale TRNs so far are available only for a limited number of model organisms, such as *Escherichia coli* (3), *Bacillus subtilis* (4), *Corynebacterium glutamicum* (5) and *Mycobacterium tuberculosis* (6). A fundamental difference between the reconstruction of TRNs and metabolic pathways is that regulatory interactions are much less conserved across bacterial genomes than metabolic pathways (7). To address the challenging problem of propagation of regulatory interactions between distant genomes, we recently developed a strategy enabling the genomic reconstruction of large-scale TRNs in diverse microbial genomes (8). The bacterial species tree is subdivided into small taxonomic families, and a subset of 5–12 representative genomes is selected for each family; then the comparative genomics techniques are used to reconstruct a collection of reference regulons in these genomes. At the next stage, these reference regulons are used for an automatic annotation of regulatory interactions in the remaining genomes from the same taxonomic group.

In 2009–10, with a goal of building collections of taxonomy-specific reference regulons, we developed two web resources for large-scale inference and analysis of regulatory interactions in prokaryotes: the RegPrecise database (http://regprecise.lbl.gov) to capture and visualize manually curated regulons (9) and the RegPredict web

*To whom correspondence should be addressed. Tel: +1 510 495 2913; Fax: +1 510 486 5614; Email: psnovichkov@lbl.org
Correspondence may also be addressed to Dmitry A. Rodionov. Tel: +1 858 646 3100; Fax: +1 858 795 5249; Email: rodionov@burnham.org

server (http://regpredict.lbl.gov) for fast and accurate regulon reconstruction (8).

The community-based approach for regulon reconstruction implemented in the RegPredict web server enabled fast accumulation of a large number of curated regulatory interactions (10–13). The current release of RegPrecise 2.0 captures the detailed descriptions of ~1000 regulogs controlled by protein transcription factors (TFs) in 13 taxonomic groups of bacteria (147 genomes total) and ~400 regulogs controlled of conserved ribonucleic acid (RNA) regulatory motifs (e.g. riboswitches) in 22 taxonomic groups (255 genomes total).

The total number of effectors of analyzed regulators in RegPrecise exceeds 200 and includes the following major classes of metabolites: amino acids, carbohydrates, nucleotides, lipids and fatty acids, co-enzymes, peptides and antibiotics, secondary metabolites and inorganic chemicals. Beside TF regulons, the last release of RegPrecise includes a large collection of curated regulons controlled by RNA regulatory elements (such as riboswitches) annotated by the RegPredict web server using RNA models from the Rfam database (14). We anticipate that the content of RegPrecise database will continue fast growing due to large-scale regulon annotation projects conducted by scientific community using RegPredict.

RegPrecise serves as a core resource on transcriptional regulation for the microbial domain of the DOE Systems Biology Knowledgebase (15), a community-driven cyber infrastructure for sharing and integrating data and analytical tools to accelerate predictive biology. In the context of this project, we developed a comprehensive set of RESTful web services for programmatic access to the whole scope of transcriptional regulatory data provided by RegPrecise.

DATA TYPES AND ORGANIZATION

The hierarchical organization of the RegPrecise database has three major levels: i) a regulon, ii) a regulog and iii) a collection of regulogs. A regulon is the basic unit of RegPrecise that represents a set of genes in a particular genome that are co-regulated by the same TF or RNA regulatory motif.

The major types of output in the RegPrecise web services application programming interface (API) are represented by six objects that provide the detailed information about reconstructed regulatory interactions (Figure 1). The Regulon object, being a basic unit, provides the general description of regulon, including the name of genome, the common name of regulator, its protein family and predicted effector molecule and the metabolic pathway (or biological process) controlled by the regulator. The regulationType property of the Regulon object identifies the type of mechanism of regulation, which is either by transcription factor (TF) or by RNA regulatory element (RNA). Three types of objects directly linked to Regulon include Regulator, Gene and Site. The Regulator object represents the actual gene encoding TF and provides its gene name, locus tag and vimssId [gene identifier in MicrobesOnline (16) database]. In rare cases, comparative genomics does not allow unambiguous selection of true cognate regulator between several homologous TFs. In this case, more than one regulator can be assigned to a particular regulon. In the case of regulation by RNA elements, no Regulator objects assigned to a regulon. The Gene object represents a regulated gene and provides general information, such as gene name, locus tag, gene vimssId and function. The Site object represents either TF-binding site or RNA regulatory element and provides its sequence, score, position relative to the translational gene start and downstream gene identifiers.

The Regulog object combines several regulons, controlled by orthologous regulators, that were reconstructed in a set of closely related genomes. Similar to the Regulon object, the general information about regulator, effector and metabolic pathway is available for the Regulog object. In addition, the taxonName property describes the name of NCBI taxon representing all genomes analyzed in a given regulog. Finally, the RegulogCollection object represents the highest level in the RegPrecise hierarchical data organization. There are six types of collections available in RegPrecise: by taxonomy, orthologous TF, TF family, RNA regulatory family, effector molecule of a regulator and regulated metabolic pathway. The collectionType property of the RegulogCollection object encodes the type of collection and can possess one of the following values: 'taxGroup', 'tf', 'tfFam', 'rnaFam', 'effector' and 'pathway'. Several types of collections, such as collection by effector and by metabolic pathway, have two-level hierarchical structure, and thus, the className property representing the upper level is provided in addition to the collection name. It should be noted that one Regulog can be assigned to several regulog collections.

WEB SERVICES API

The developed web services API enables programmatic access to the whole content of the RegPrecise database. The API is implemented as a set of RESTful web services providing data in either JavaScript Object Notation (JSON) or Extensible Markup Language (XML), two of the most popular formats. The base Uniform Resource Locator (URL) for all web services is http://regprecise.lbl.gov/Services/rest/.

The current version of RegPrecise web services API includes 14 resources that can be classified into four categories (Table 1). The core resources provide access to the regulators (genes encoding either TFs or RNA motifs), target regulated genes and sites (either TF-binding sites or RNA regulatory elements). These three resources together provide complete information about the content of a particular regulon and can be queried by regulon identifier. At the same time, if a user is interested in all genes regulated by orthologs of a particular TF in a group of closely related genomes, the *genes* resource can be queried by an identifier of a corresponding regulog. The same is true for two other resources. The *regulog* and *regulon* resources can be used to obtain summary

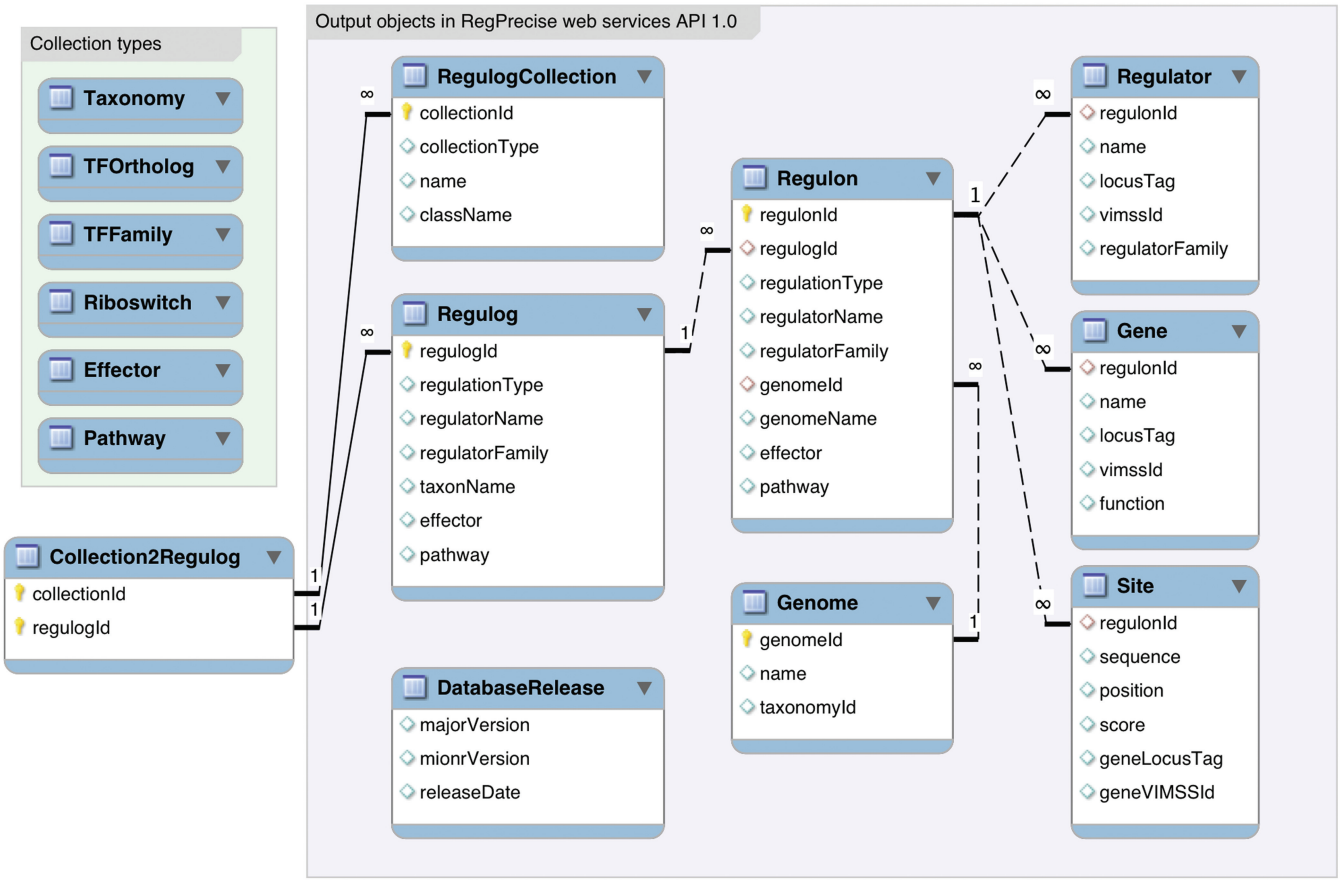

Figure 1. The major output objects and their properties in RegPrecise web services API.

Table 1. List of resources available in RegPrecise web services API 1.0

Resources	Description
Core resources	
/regulators	Represents a list of analyzed regulators that belong to either a given regulon or regulog.
/genes	Represents a list of regulated genes that belong to either a given regulon or regulog.
/sites	Represents a list of TF-binding sites or RNA regulatory elements that belong to either a given regulon or regulog.
/regulog	Represents a regulog.
/regulon	Represents a regulon.
Navigation resources	
/regulogCollections	Represents various types of regulog classifications: by taxonomic group, orthologous TF, TF family, RNA regulatory element, metabolic pathway or biological process, effector molecule or environmental signal.
/genomes	Represents a list of genomes that have at least one reconstructed regulon.
/regulogs	Represents a list of regulogs in a collection of a given type and identifier.
/regulons	Represents a list of regulons in either a particular regulog or a genome.
Statistics resources	
/regulogCollectionStats	Represents general statistics on regulog collections of a particular type.
/genomeStats	Represents general statistics on genomes that have at least one reconstructed regulon.
Utility resources	
/searchRegulons	Search regulons by the name or locus tag of either regulator or target regulated genes.
/searchExtRegulons	Search regulons by a genome and list of gene locus tags.
/release	Represents the version and the release date of the current version of RegPrecise.

information about the intrinsic attributes of a regulog/regulon, such as the associated metabolic pathway (or biological process), effector molecules, the regulation mode (activation or repression), the regulation type (TF or RNA) and regulator family.

The navigation resources allow traversing through the hierarchical organization of the RegPrecise data (Table 1). The list of the available regulog collections can be retrieved for each of these types using the *regulogCollections* resource. Subsequently, the list of

regulogs that belong to a particular collection can be obtained by the *regulogs* resource. It is important that the collection identifiers are unique only within the space of regulog collections of a particular type. Thus in addition to these identifiers, it is necessary to provide the type of a collection as a parameter in a query. Finally, a list of regulons analyzed for a particular TF of RNA motif in a set of closely related genomes can be accessed by the *regulons* resource given regulog identifier. The alternative navigation route starts from a list of genomes that have at least one reconstructed regulon. The complete list of genomes can be obtained by the *genomes* resource. The same *regulons* resource can be used to get a list of all reconstructed regulons for a particular genome.

To get an overview of the RegPrecise content, we developed two resources providing the general statistics for regulog collections (*regulogCollectionStats* resource) and genomes (*genomeStats* resource). The statistics includes the number of genomes, reconstructed regulogs, regulons, and regulatory sites for the TFs and RNA elements.

Finally, the utility resources allow for searching for regulons by locus tags of the target regulated genes or TFs (*searchRegulons* resource). We also developed a special resource that can be used to analyze gene sets. The *searchExtRegulons* resource takes NCBI taxonomy id of a genome and comma-separated list of gene locus tags and returns the non-redundant list of regulons that contain at least one of the provided genes. In particular, this resource can be used for automatic validation of gene clusters reconstructed by the analysis of expression data. The current version of the RegPrecise database underlying the web services API can be obtained by the *release* resource.

The complete documentation on the listed resources can be found at http://regprecise.lbl.gov/RegPrecise/services.jsp. Two types of client code examples are provided. First, we provide a template program in perl that can be run to access several of the web services and parse the output data. The program is organized in two perl scripts: (i) RegPreciseAdapter.pm—a perl module that provides access to the individual web services and (ii) regulons.pl—an example of workflow that can be implemented using a combination of several web services. Both scripts can be easily modified to access all the web services available. At the same time, each web service is accompanied with an example of accessing API using cURL command-line tool.

USE CASES

In this study, we provide two examples of the possible scenarios of using RegPrecise web services API.

Scenario 1—obtaining information on a TRN available for a particular genome. To check the availability of regulatory interactions for a given genome, the user can query the *genomes* resource. By analyzing the output for the presence of the NCBI taxonomy id of the genome of interest, the user can confirm the presence of this genome in RegPrecise. For instance, the NCBI taxonomy id of *Shewanella baltica* OS155 is 325240, which corresponds to the internal genome identifier genomeId = 7. The list of all reconstructed regulons in this genome can be retrieved by querying *regulons?genomeId = 7*. Analysis of the *regulationType* property of each regulon in the output will show that 63 regulons are regulated by TFs, whereas 14 regulons are controlled by RNA regulatory elements. Among TF-operated regulons, the tryptophane regulon TrpR has regulonId = 6378. The list of TrpR-regulated genes can be retrieved by querying *genes?regulonId = 6378*.

Scenario 2—obtaining all genes that are regulated by a particular RNA motif in any genome. For example, the user is interested in genes regulated by FMN riboswitch. Analysis of the *regulogCollections?collectionType = rnaFam* query output identifies the corresponding regulog collection (collectionId = 25). List of all reconstructed regulogs in this collection can be retrieved by querying *regulogs?collectionType = rnaFam&collectionId = 25*.

Finally, by iterating through each regulog, the complete list of genes regulated by FMN riboswitch can be retrieved by *genes* resource by providing regulog identifier as a parameter, e.g. *genes?regulogId = 1450*.

FUTURE DEVELOPMENT

We are currently working on the automatic conservative propagation of all regulons inferred in the reference set of genomes to all closely related genomes from the same taxonomic group. We will develop new web services to enable programmatic access to the results of propagation.

ACKNOWLEDGEMENTS

The authors are grateful to Andrei Osterman, Alexey Kazakov and KBase team for testing and useful discussions.

FUNDING

This work was supported by the Office of Science, Office of Biological and Environmental Research, of the US Department of Energy (DOE) under contract No. [DE-AC02-05CH11231], as part of the DOE Systems Biology Knowledgebase, and under contract No. [DE-SC0004999] with Sanford-Burnham Medical Research Institute and Lawrence Berkeley National Laboratory, as part of Genomic Science Program. Funding for open access charge: DOE contract No. [DE-AC02-05CH11231].

Conflict of interest statement. None declared.

REFERENCES

1. Caspi,R., Altman,T., Dreher,K., Fulcher,C.A., Subhraveti,P., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Mueller,L.A. *et al*. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*., **40**, D742–D753.

2. Henry,C.S., Dejongh,M., Best,A.A., Frybarger,P.M., Linsay,B. and Stevens,R.L. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.*, **28**, 977–982.
3. Gama-Castro,S. and Jiménez-Jacinto,V. (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36**, D120–D124.
4. Sierro,N., Makita,Y., De Hoon,M. and Nakai,K. (2007) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res.*, **36**, D93–D96.
5. Baumbach,J. (2007) CoryneRegNet 4.0—a reference database for corynebacterial gene regulatory networks. *BMC Bioinformatics*, **8**, 429.
6. Jacques,P.E., Gervais,A.L., Cantin,M., Lucier,J.F., Dallaire,G., Drouin,G., Gaudreau,L., Goulet,J. and Brzezinski,R. (2005) MtbRegList, a database dedicated to the analysis of transcriptional regulation in *Mycobacterium tuberculosis*. *Bioinformatics*, **21**, 2563–2565.
7. Rodionov,D.A. (2007) Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chem. Rev.*, **107**, 3467–3497.
8. Novichkov,P.S., Rodionov,D.A., Stavrovskaya,E.D., Novichkova,E.S., Kazakov,A.E., Gelfand,M.S., Arkin,A.P., Mironov,A.A. and Dubchak,I. (2010) RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic Acids Res.*, **38**, W299–W307.
9. Novichkov,P.S., Laikova,O.N., Novichkova,E.S., Gelfand,M.S., Arkin,A.P., Dubchak,I. and Rodionov,D.A. (2010) RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Res.*, **38**, D111–D118.
10. Rodionov,D.A., Novichkov,P.S., Stavrovskaya,E.D., Rodionova,I.A., Li,X., Kazanov,M.D., Ravcheev,D.A., Gerasimova,A.V., Kazakov,A.E., Kovaleva,G.Y. *et al.* (2011) Comparative genomic reconstruction of transcriptional networks controlling central metabolism in the Shewanella genus. *BMC Genomics*, **12(Suppl. 1)**, S3.
11. Ravcheev,D.A., Best,A.A., Tintle,N., Dejongh,M., Osterman,A.L., Novichkov,P.S. and Rodionov,D.A. (2011) Inference of the transcriptional regulatory network in Staphylococcus aureus by integration of experimental and genomics-based evidence. *J. Bacteriol.*, **193**, 3228–3240.
12. Ravcheev,D.A., Li,X., Latif,H., Zengler,K., Leyn,S.A., Korostelev,Y.D., Kazakov,A.E., Novichkov,P.S., Osterman,A.L. and Rodionov,D.A. (2012) Transcriptional regulation of central carbon and energy metabolism in bacteria by redox-responsive repressor rex. *J. Bacteriol.*, **194**, 1145–1157.
13. Leyn,S.A., Li,X., Zheng,Q., Novichkov,P.S., Reed,S., Romine,M.F., Fredrickson,J.K., Yang,C., Osterman,A.L. and Rodionov,D.A. (2011) Control of proteobacterial central carbon metabolism by the HexR transcriptional regulator: a case study in Shewanella oneidensis. *J. Biol. Chem.*, **286**, 35782–35794.
14. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al.* (2010) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, **39**, D141–D145.
15. DOE Systems Biology Knowledgebase http://genomicscience.energy.gov/compbio/. (6 January 2012, date last accessed).
16. Dehal,P.S., Joachimiak,M.P., Price,M.N., Bates,J.T., Baumohl,J.K., Chivian,D., Friedland,G.D., Huang,K.H., Keller,K., Novichkov,P.S. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.*, **38**, D396–D400.